A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans

Ana I. Vazquez,*,[1] Gustavo de los Campos,* Yann C. Klimentidis,* Guilherme J. M. Rosa,† Daniel Gianola,† Nengjun Yi,* and David B. Allison*

*Section on Statistical Genetics, Department of Biostatistics, University of Alabama, Birmingham, Alabama 35294, and †Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53705

ABSTRACT Prediction of genetic risk for disease is needed for preventive and personalized medicine. Genome-wide association studies have found unprecedented numbers of variants associated with complex human traits and diseases. However, these variants explain only a small proportion of genetic risk. Mounting evidence suggests that many traits, relevant to public health, are affected by large numbers of small-effect genes and that prediction of genetic risk to those traits and diseases could be improved by incorporating large numbers of markers into whole-genome prediction (WGP) models. We developed a WGP model incorporating thousands of markers for prediction of skin cancer risk in humans. We also considered other ways of incorporating genetic information into prediction models, such as family history or ancestry (using principal components, PCs, of informative markers). Prediction accuracy was evaluated using the area under the receiver operating characteristic curve (AUC) estimated in a cross-validation. Incorporation of genetic information (*i.e.*, familial relationships, PCs, or WGP) yielded a significant increase in prediction accuracy: from an AUC of 0.53 for a baseline model that accounted for nongenetic covariates to AUCs of 0.58 (pedigree), 0.62 (PCs), and 0.64 (WGP). In summary, prediction of skin cancer risk could be improved by considering genetic information and using a large number of single-nucleotide polymorphisms (SNPs) in a WGP model, which allows for the detection of patterns of genetic risk that are above and beyond those that can be captured using family history. We discuss avenues for improving prediction accuracy and speculate on the possible use of WGP to prospectively identify individuals at high risk.

SKIN cancer is the most common form of cancer, and its incidence has increased in recent decades. In Queensland, Australia, it has been estimated that half of the population is likely to develop skin cancer during their lifetime (World Cancer Report 2008). In its most severe form (*i.e.*, melanoma), skin cancer can be deadly. Although protection against sunburn (Ziegler *et al.* 1994) is widely believed to reduce the harmful effects of sun exposure on the skin, many individuals continue to seek out intense sun exposure without such protection (Robinson 1990). This may be due in part to belief that their risk of skin cancer is too low to be a serious concern. Such beliefs can be maintained and rationalized by the observation that many individuals exposed to such risk do not experience the adverse event. If individuals could be provided with personalized information about their individual risk, it might promote greater use of preventive measures among those at greatest risk.

Although ultraviolet (UV) exposure and light skin pigmentation are major risk factors for all types of skin cancers [*e.g.*, it is estimated that 80% of melanoma is caused by ultraviolet damage to sensitive skin (IARC 1992)], evidence suggests that genetic factors can also play a role, independent of skin pigmentation. Predictive models are usually based on standard covariables and family history. Additionally, several genetic variants, such as the MC1R, ASIP, TYR, EXOC2, and UBAC2 and the 1p36 and 1q42 loci, have been shown to be associated with basal and squamous cell carcinomas, as well as with melanomas, independent of skin pigmentation (Gudbjartsson *et al.* 2008; Stacey *et al.* 2008). These variants typically account for a small proportion of genetic-based disease risk (Han *et al.* 2006; Pharoah 2008).

As with other phenotypic and disease traits, the inability of loci discovered by genome-wide association studies (GWAS)

Copyright © 2012 by the Genetics Society of America
doi: 10.1534/genetics.112.141705
Manuscript received May 3, 2012; accepted for publication September 7, 2012
Available freely online through the author-supported open access option.
Reference numbers: dbGaP Study Accession: phs000007.v1.p1, IRB protocol number: X100712003, IRB protocol title: "Prediction of Diseases Phenotypes Using Whole Genome Markers from the Framingham Population Data Set".
[1] Corresponding author: Section on Statistical Genetics, Department of Biostatistics, University of Alabama, 1665 University Blvd., Ryals Public Health Bldg. 414, Birmingham, AL 35294. E-mail: anainesvs@gmail.com.

to explain a substantial proportion of heritability has led to much debate regarding where this so-called "missing heritability" lies (Manolio *et al.* 2009). It has been suggested that the underlying genetic architecture of many human traits and diseases may involve a substantial number of small-effect genes, thus conforming to the so-called infinitesimal model of quantitative genetics (Fisher 1918; Bulmer 1980; Lander and Schork 1994; Goddard and Hayes 2007). However, in most genetic risk prediction models currently being tested, only a few [*i.e.*, <500 single-nucleotide polymorphisms (SNPs)] statistically significant SNPs are included.

The recognition that complex human traits and diseases could be affected by a large number of genes has motivated many researchers in other fields (Lee *et al.* 2008; Wray *et al.* 2008; Purcell *et al.* 2009; Hill 2010; Yang *et al.* 2010; de los Campos *et al.* 2010a) to propose the use of statistical methods tailored for the prediction of complex traits. These methods, largely developed in the field of animal breeding, were first proposed by Meuwissen *et al.* (2001), who suggested predicting genetic factors by regressing phenotypes on a large number of markers covering the entire genome. The markers are assumed to be in linkage disequilibrium (LD) with one or many loci affecting the phenotypic traits, and the estimates of individual marker effects are expected to be small (Goddard and Hayes 2007). Such models and variations thereof have been used successfully in animal and plant breeding for prediction of production-related traits (de los Campos *et al.* 2009; Hayes *et al.* 2009; VanRaden *et al.* 2009; Crossa *et al.* 2010; Weigel *et al.* 2010). More recently, several authors have proposed and used this methodology for the prediction of complex human traits such as height (Yang *et al.* 2010; Makowsky *et al.* 2011) and several cancer outcomes (Vazquez 2010).

In this study, we determine whether genetic predisposition to skin cancer could be used to predict disease outcome. Compared with height, skin cancer is less heritable, more complex, and highly relevant. Due to these features, we compared different methods to account for genetic susceptibility to skin cancer: (1) pedigree-based and SNP-based predictions via (2) whole-genome prediction (WGP) for liability to skin cancer, using thousands of evenly distributed markers across the genome and via (3) the principal components of a subset of independent SNPs (previously used to predict geographical origin). To perform this study, we extended the Bayesian LASSO (Park and Casella 2008) regression with a probit link (Dempster and Lerner 1950) to model skin cancer.

Materials and Methods

Data

The data set consists of 5132 participants from the Framingham Heart Study, which has collected phenotypic information across three generations of families (Dawber *et al.* 1951, 1963). Subjects in this study have been characterized every other year from adulthood to death on risk factors, outcomes of physical exams, and disease status. Participants included in our study belong to the original cohort (n = 1498) and to the offspring cohort (n = 3634), with a total of 2319 males and 2813 females, all of whom were genotyped. Subjects from the third generation cohort were not included in our study because the follow-up period of this cohort was too short. The skin cancer outcomes were collected by Bernard E. Kreger (Boston University, Boston, study accession no. pht000039) (Kreger *et al.* 1991). The study declares cancerous all subjects whose pathology reports confirm their cancer. After the 1980s, cases were validated using medical records. The available data represent primary tumors only. The skin cancer outcomes study was updated in 2006, containing life-long follow up, *i. e.*, 1948–2006 for the original cohort and 1971–2006 for the offspring cohort.

All subjects were genotyped for SNPs with the Affymetrix 500K chip. Due to computational limitations (memory requirements), we fitted models using 41,188 evenly spaced SNPs. Evidence from U.S. Holstein cattle has indicated that predictive ability for several traits does not increase markedly when using >10,000 SNPs for a panel of 50,000 maximum (Vazquez *et al.* 2010). Nevertheless, the degree of linkage disequilibrium differs across species. However, a recent study on human height with the same data set indicated that, in this population and with this sample size (n = 5117), the increase in predictive ability when using >30,000 SNPs was limited (Makowsky *et al.* 2011).

Statistical models

Full model: The outcome (y_i) was defined as presence (y_i = 1) or absence (y_i = 0) of skin cancer at any site, excluding skin of labia majora, vulva, penis, and scrotum, as a primary tumor site. We modeled probability of skin cancer using the probit link or threshold model (Dempster and Lerner 1950; Harville and Mee 1984). Here, probability of disease equals the standard normal cumulative density function, $\Phi(.)$, evaluated at a subject-specific risk score, with either model

$$\eta_i = \beta_0 + \sum_{j=1}^{p_1} x_{1ij}\beta_{1j} + \sum_{j=1}^{p_2} x_{2ij}\beta_{2j}$$

or model

$$\eta_i = \beta_0 + \sum_{j=1}^{p_1} x_{1ij}\beta_{1j} + u_i,$$

which was represented as the sum of an intercept (β_0) plus a regression on the "fixed effects" of sex (as dummy variable), cohort (a factor with three levels), and ethnicity covariates (explained below) ($\sum_{j=1}^{p_1} x_{1ij}\beta_{1j}$), plus either a "random effects" regression on marker genotypes ($\sum_{j=1}^{p_2} x_{2ij}\beta_{2j}$) or a random effects u_i being the genetic liability to skin cancer derived based on pedigree connections. Therefore, the joint conditional probability of the data, $\mathbf{y} = \{y_i\}$, given the unknown regression coefficients, $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2\} = \{\mathrm{\beta}_0, \mathrm{\beta}_{11}, ..., \beta_{1p_1}, \beta_{21}, ..., \beta_{2p_2}\}$ and $\mathbf{u} = \{u_1, ..., u_{5132}\}$, was

$$p(\mathbf{y}|\boldsymbol{\beta},\mathbf{u}) = \prod_{i=1}^{5132}\left\{[\Phi(\eta_i)]^{y_i}[1-\Phi(\eta_i)]^{1-y_i}\right\}.$$

Assigning a prior density to the vector of model unknowns, $\boldsymbol{\beta}$, and $\mathbf{u}$, completes the Bayesian model. We assigned a flat prior to the intercept and to the effects of sex, cohort, and ethnicity covariates. This yielded estimates of effects comparable to those obtained with maximum likelihood and used the Bayesian Lasso of Park and Casella (2008) to structure the prior density of marker effects. This prior density yielded shrunken estimates of marker effects and has been successfully used for WGP (de los Campos *et al.* 2009). The joint prior density was

$$\begin{aligned} p(\beta_0,\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\mathbf{u},\boldsymbol{\tau}^2,\lambda) \propto & \left[\prod_{j=1}^{p} N\left(\beta_{2,j}\middle|0,\tau_j^2\right)\right] \\ & \times \left[\prod_{j=1}^{p}\mathrm{Exp}\left(\tau_j^2\middle|\lambda^2\right)\right] \times G(\lambda^2|\alpha_1,\alpha_2) \\ & \times N\left(\mathbf{u}\middle|\mathbf{0},\mathbf{A}\sigma_u^2\right)\times\chi^{-2}\left(\sigma_u^2|S,df\right), \end{aligned}$$

where $N(\beta_{2,j}|0,\tau_j^2)$ is a normal density assigned to the j^{th} marker effect; $j \in (1,...,p_2)$ with prior mean equal to zero and marker-specific prior variance (τ_j^2); $\mathrm{Exp}(\tau_j^2|\lambda^2)$ is an exponential prior for variances of marker effects τ_j^2; $\mathrm{G}(\lambda^2|\alpha_1,\alpha_2)$ is a gamma distribution for λ^2 with shape and rate parameters α_1 and α_2, respectively; $N(\mathbf{u}|\,\mathbf{0},\mathbf{A}\sigma_u^2)$ is a normal distribution for $\mathbf{u}$ with mean $\mathbf{0} = \{0_1,...,0_{5132}\}$ and variance $\sigma_u^2 \times \mathbf{A}$, which is the additive genetic relationship matrix based on the pedigree; and finally $\chi^{-2}(\sigma_u^2|S,df)$ is an inverse chi-square distribution for σ_u^2 with parameters S and df. In the analysis, these hyperparameter values were set to $\alpha_1 = 1.5$ and $\alpha_2 = 1e-4$; this gives a relatively flat prior over a wide range of values of the regularization parameter (see Pérez *et al.* 2010 for further details), as S and df were 0.19 and 5, respectively. Models were fitted using a modified version of the BLR package (de los Campos and Pérez 2010) of R (R Development Core Team 2010), which can be used for regressions for binary outcomes according to the model described above.

Sequence of models: Using the specification described above, we defined a sequence of models and evaluated the performance of each model, using cross-validation. Our baseline model (covariates) included only the effects of sex and cohort. This model was first extended by adding to the regression the effects of the first two principal components of a set of 1000 European ethnicity-informative SNPs previously reported (model denoted as PC-SNP). The panel of ethnicity-informative SNPs used here was those reported by Drineas *et al.* (2010). Figure 1 shows a scree plot of the first 20 eigenvalues derived from the markers reported by Drineas *et al.* In a number of studies, the first two principal components (PC1 and PC2) of a large set of genetic variants have been shown to be effective predictors of the ancestral/geographical origin (latitude and longitude) of individuals of European descent (Price *et al.* 2008; Tian *et al.* 2008; Novembre *et al.* 2008; Drineas *et al.* 2010). For this reason we included two PCs in the model, even though PC2 is only slightly higher than the following PCs. Additionally, the covariates model was extended by adding subsets of SNPs (250, 500, 1000, 50,000, 10,000, and 41,000) distributed over the whole human genome; these models are denoted as 0.25K-SNP, 0.5K-SNP, 1K-SNP, 5K-SNP, 10K-SNP, and 41K-SNP, respectively. The 41,000 SNPs were obtained by choosing ~1 of every 12 SNPs from the original SNP panel. The SNPs in the smaller sets are all included in the larger sets. In this series of increasingly denser models, we aim to discover how many markers are needed to increase the predictive ability. Finally, we extended the covariates model by adding a random effect representing a regression on the pedigree, and this model was denoted as pedigree.

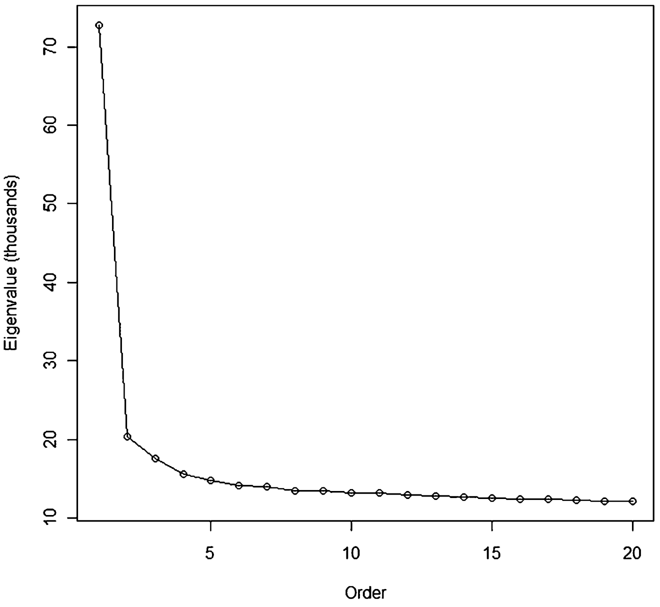

Figure 1 First 20 eigenvalues derived from ethnicity-informative panel of 1000 SNPs.

Estimated probabilities and odds ratio: Probabilities and odds ratio for the relative risk of developing skin carcinoma for groups of sex, cohort, and ethnicity were estimated with the PC-SNP model at a fixed level of the dichotomous covariables and fixing the principal components at the mean, first, and third percentile values of the eigenvectors of PC1 and PC2 (see Table 2). Similarly, probabilities and odds ratio of the genetic effects $(\sum_{j=1}^{p_2}x_{2ij}\beta_{2j})$ were estimated with the 41K-SNP model for differences in the first and third percentiles of the genetic effects for fixed levels of sex and cohort.

Assessment of model prediction performance: Models were compared based on prediction accuracy, evaluated in a 20-fold cross-validation with subjects assigned to folds at random. The 20-fold cross-validation yielded predictions of risk scores $\{\hat{\eta}_i\}$

that were derived without using the i^{th} observation or any of the observations assigned to the same fold to which the i^{th} observation was assigned. Using the pairs of points $\{y_i, \hat{\eta}_i\}$, we estimated false positive rate and area under the receiver operating characteristic curve (AUC) (see Fawcett 2006), using the R package ROCR (Sing *et al.* 2005).

To assess the uncertainty of the AUC estimate due to sampling variability, we performed 500 random partitions in training and testing sets with sizes equal to those of the cross-validation folds (testing $n = 257$ subjects and training $n = 4875$ subjects), maintaining the sizes of the subsets in the cross-validation (*i.e.*, on each replicate, 5% of the individuals were randomly assigned to testing and the remaining 95% to training). Each replicate yielded an estimate of AUC by model, and variability across replicates was reflective of uncertainty due to sampling of training and testing data sets. From this analysis, we reported the number of times one model outperformed another for AUC.

Results

Descriptive statistics and parameter estimates

The skin cancer incidence in our data set was 14.1% for the entire period evaluated, starting at 1948 with the original cohort and at 1971 with the offspring cohort, and followed until the 2006 update. Incidence did vary, however, across sex and cohorts. The incidence was higher in males (16%) than in females (13%) and higher in the original cohort (17%), which had a longer follow-up period, than in the offspring cohort (13%). The first two eivenvectors of the PCs decomposition derived from 1000 ethnicity-informative SNPs are displayed in Figure 2A. Figure 2B shows the empirical distribution of PC1 for individuals with and without skin cancer. PC1 has the highest discriminating power (Figure 2A) and, at the marginal level, lower values at the eigenvector were associated with higher incidence of skin cancer (Figure 2B). This PC1 has been reported to track northern *vs.* southern European ancestry (Drineas *et al.* 2010). Further evidence of the marginal association of incidence of skin cancer and PC1 is given in Table 1, where average incidence of skin cancer is presented by quartiles of PC1 and PC2.

In the PC-SNP model, the effects of cohort, sex, and the first two PCs are estimated jointly. We found that the estimated coefficient for male sex in the liability scale was 0.18 [0.09, 0.26] (posterior mean and 95% credibility region in brackets). This estimate implies an increased higher risk of developing skin cancer in males, relative to females. The estimated coefficient for the original cohort with respect to the offspring cohort as baseline was 0.32 [0.23, 0.42], also in liability scale. This implies higher risk of developing skin cancer for members of the original cohort and likely reflects the effect of a longer follow-up period for members of this cohort (original cohort started 23 years before the offspring cohort). The estimated coefficients for the first and second PCs were −11.19 [−14.47, −7.87] for PC1 and 12.97 [9.41, 16.62] for PC2, indicating that risk increases as PC1 decreases and as PC2 increases. All 95% confidence regions for the estimated effects did not include zero, showing evidence of nonnull effects of the predictors considered on skin cancer risk. All the estimates presented above are in the scale of the linear predictor (or liability scale). Given the nonlinearity of the model, these results are difficult to interpret. Table 2 shows the estimated probability of skin cancer risk (and estimates of 95% posterior credibility regions) for different combinations of the predictor variables using the PC-SNP model. The 95% credibility regions of the probability estimates for the two cohorts and for gender do not overlap, suggesting significant differences for these predictors. The odds ratio for the cohort variable is 1.76 [1.50, 2.10] (original relative to offspring cohort) in males and 1.81 [1.52, 2.18] in females, while the odds ratio for sex is 1.36 [1.17, 1.58] (male relative to female) in the original cohort and 1.39 [1.19, 1.64] in the offspring cohort. All were estimated at the mean value of the two PCs.

When the covariates model was extended by adding genetic effects connected by the pedigree, the predicted genetic effects $(\hat{u}_i)$ in the liability scale ranged between −0.5 and 1.37 (Figure 3). Likewise, the covariates model was also extended by adding the joint effects of SNPs evenly spaced along the genome (from 250,000 to 41,000). In models including genome-wide SNPs, the total contribution of all SNPs, to the cancer risk, is summarized by the linear score $\hat{g}_i = \sum_{j=1}^{p_2} x_{2ij}\hat{\beta}_{2j}$, where x_{2ij} are marker genotypes and $\hat{\beta}_{2j}$ are estimates of marker effects. In our sample, this score $(\hat{g}_i)$ ranged between −1.07 and 2.7 for the model with 41,000 markers (Figure 3). These results suggest the existence of variation due to genetic factors that was captured by either the pedigree or the markers. Figure 3 shows the predicted genetic scores derived from the pedigree $(\hat{u}_i)$ and the whole genome regression (WGR) (41K-SNP) $(\hat{g}_i)$ (both derived from models fitted to the entire data set). The correlation between the genetic scores was 0.783. Both scores exhibit a bimodal distribution: the group with lower scores corresponds to individuals with no personal or family history of skin cancer, while the second group primarily includes individuals with some personal or family history of skin cancer. The within-group dispersion of the pedigree-based score around the mean of the clusters is much smaller than that of the score derived from the WGR. This occurs because, although WGR captures family history, this approach also allows for the borrowing of information across nominally unrelated individuals.

Evaluation of the models' predictive performances

The estimates presented in the preceding section indicate that all the predictor variables are significantly associated with the risk of developing skin cancer. Additionally, we evaluated the prediction accuracy of each of the models with a 20-fold cross-validation to assess how useful each of these models is in the assessment of risk in individuals with yet-to-be-observed skin cancer outcomes. Figure 4 shows the AUC obtained in the 20-fold cross-validation for (Figure 4A) the

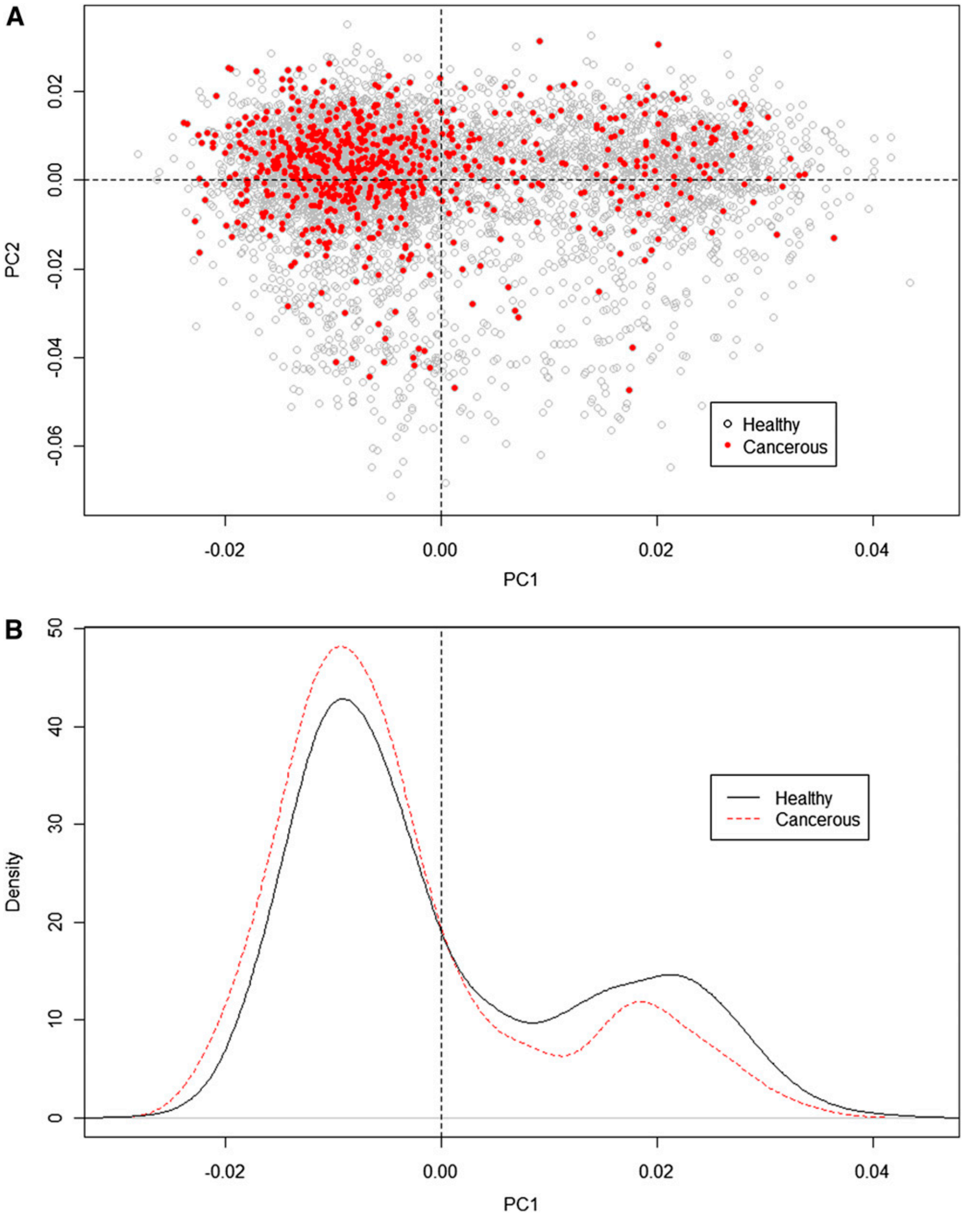

Figure 2 (A) First (*x*-axis) and second (*y*-axis) principal components eigenvectors derived from 1000 ethnicity-informative SNPs (red dots correspond to subjects that developed skin cancer, and gray dots correspond to healthy subjects). (B) Empirical distribution of the first principal component separated by cancerous or healthy subjects.

covariates, pedigree, PC-SNP, and 41K-SNP, and for (Figure 4B) genomic-enabled models at increasing SNP density from zero to 41,000. The covariates model had an AUC of 0.534 (baseline model). Accounting also for the pedigree relations yielded an AUC of 0.579, improving the AUC by 8.4% [calculated as $100\times(0.579-0.534)/0.534$]. The PC-SNP had an AUC of 0.622, 16.5% higher than that of the baseline model. Finally, the 41K-SNP model had an AUC of 0.635, 18.9% higher than that of the baseline model. The evaluation of the AUC of the models including sex, cohort, and varying numbers of evenly spaced SNPs showed a monotonic increase in AUC with the number of SNPs, from the covariates model (with zero SNPs) to 41K-SNP (Figure 4B).Family relationships between training and testing sets have been shown to affect prediction accuracy (Habier *et al.* 2010; Pérez-Cabal *et al.* 2012). To assess the impact of family relationships on

Table 1 Incidence of skin cancer by levels defined using the first and second eigenvectors of the ethnicity SNP-derived principal components

	Group			
	$v_i \leq q_{0.25}$	$q_{0.25} < v_i \leq q_{0.50}$	$q_{0.50} < v_i \leq q_{0.75}$	$v_i > q_{0.75}$
First principal component	0.178	0.150	0.126	0.098
Second principal component	0.100	0.150	0.152	0.163

v_i, Value of the first and second principal component in subject i; $q_.$, corresponding quartile.

Table 2 Estimated probabilities and 95% credibility region (CR) of developing skin cancer for different levels of the predictor variables, derived from a model including sex, cohort, and the first two principal components of 1000 ethinicity-informative SNPs

		Probability of developing skin cancer	
Cohort	Sex	Estimate	CR 95%
Original	Male	0.242	[0.213, 0.273]
Offspring	Male	0.153	[0.138, 0.171]
Original	Female	0.190	[0.167, 0.215]
Offspring	Female	0.115	[0.103, 0.129]

the prediction accuracy derived from models incorporating pedigree or markers, we calculated the AUC of each of the models for subjects that did not have data from relatives in the training data sets (n = 871) and for those that had at least one relative in the training data set (n= 4248) (see Table 3). Results show that in the pedigree and the 41K-SNP model, part of the gains in prediction accuracy can be explained by information provided by relatives. However, as shown in Figure 4A, the 41K-SNP model outperformed the pedigree model, and in Figure 4B the 41K-SNP model had higher predictive accuracy than all the other models for individuals whose risk was predicted without having any relative in the training data set (first row in Table 3). Therefore, we conclude that although part of the prediction accuracy of the 41K-SNP model can be explained by information coming from relatives, this model is capturing patterns of genetic risk beyond those that can be captured by family history.

Assessing the sampling variability of the AUC estimates: The above results suggest that whole-genome markers can increase prediction accuracy of skin cancer susceptibility by a nonnegligible margin. To evaluate uncertainty about these point estimates, we replicated a training–testing evaluation 500 times (see *Materials and Methods* above). The covariates model was improved by the pedigree model in 70% of the replicates, by the PC-SNP model in 90% of the replicates, and by the 41K-SNP model in 94% of the replicates. The 41K-SNP model had higher prediction accuracy than the pedigree model in 90% of the replicates and was higher in accuracy than the PC-SNP model in 66% of the replicates. Figure 5 shows predictive correlation and AUC results by replicate for two models simultaneously (y-axis and x-axis). At the 45° line, both models performed equally, while above the line, the model represented on the y-axis performed better and vice versa.

Discussion

Factors affecting skin cancer

Skin cancer is the most frequent form of cancer. Further, nonmelanoma skin cancer is the most frequent type of cancer in light-skinned populations (World Cancer Report 2008). In our sample, skin cancer was also the most prevalent type of cancer. The main risk factors for skin carcinogenesis are ultraviolet light exposure, skin type, and geographical location [*e.g.*, no-melanoma skin cancer is 5 times higher in the United

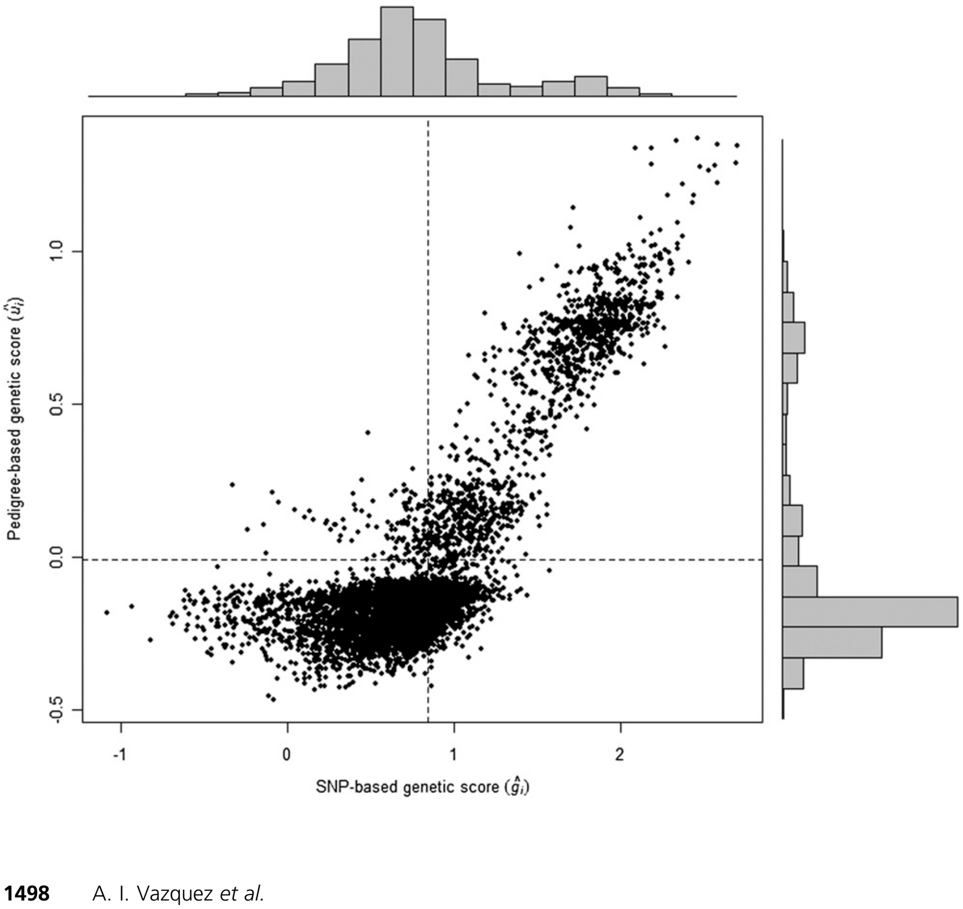

Figure 3 Scatter plot of the pedigree-based predicted genetic risk for skin cancer and the SNP-based ones ($\hat{u}_i$ and $\hat{g}_i$, respectively), as well as the histogram of their distribution.

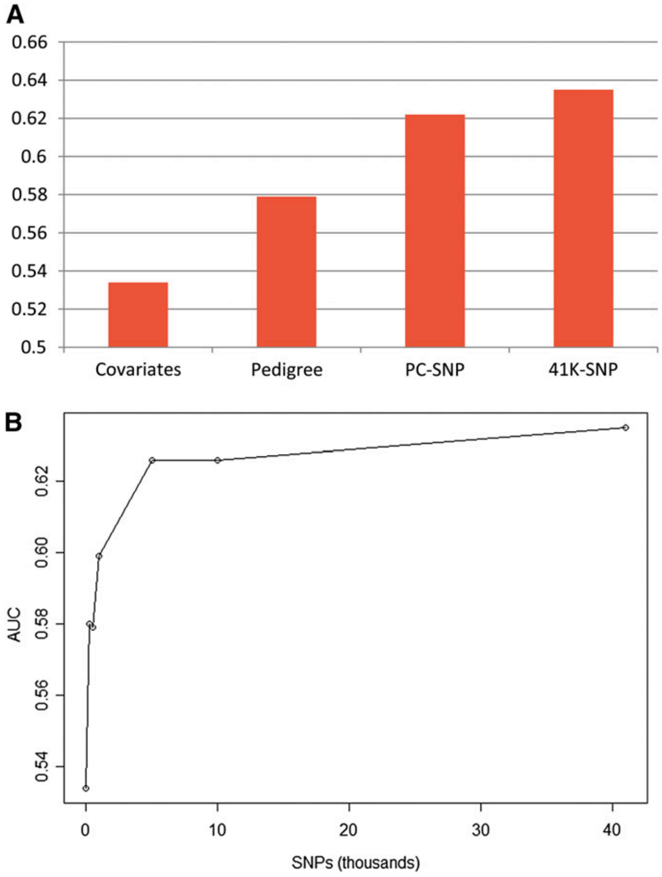

Figure 4 Mean area under the curve for 20-fold cross-validation for (A) a model without any genetic information and two models with genetic information, one including pedigree and a WGP model, and (B) for WGP models of increasing number of SNPs.

States and 20–40 times higher in Australia than in Europe (World Cancer Report 2008)]. In our study, geographical location was not a source of variation, since all data came from the same geographical location (Framingham, MA). However, the harm from sun exposure may have increased over time due to, *e.g.*, atmospheric changes. Indeed, it has been estimated that incidence of nonmelanoma skin cancer increased by 77% from 1992 to 2006 (Stern 2010, p. 13). The estimated effect of cohort in our data might be simultaneously reflecting two factors that may have opposing effects: subjects from the original cohort have had longer exposure (offspring cohort data collection started 23 years later) and year of exposure (since skin cancer incidence has increased over the past 50 years). Gender differences have also been reported in the literature, and they indicate a lower incidence among females (Diepgen and Mahler 2002). Our results are consistent with this (Table 2). The association between sex and risk of developing skin cancer has been largely attributed to different lifestyles (*e.g.*, males are more exposed to sun and less likely to use sun protection) (McCarthy *et al.* 1999). Additionally, there is evidence of sex-based biological differences at the skin level, relevant to skin cancer liability (Thomas-Ahner *et al.* 2007).

Table 3 Area under the curve estimated in the subjects that have no relatives in the training set and in the subjects that do, for all the models

	Covariates	Pedigree	PC-SNP	41K-SNP
No relatives in training set	0.540	0.549	0.635	0.629
At least one relative in training set	0.531	0.583	0.619	0.637

Prediction of genetic risk to skin cancer

Previous studies (Han *et al.* 2006; Gudbjartsson *et al.* 2008; Stacey *et al.* 2008; World Cancer Report 2008) have indicated that genetic factors play an important role in predisposition to skin cancer. However, predictive models for skin cancer, although accurate, do not usually account for genetic factors (*e.g.*, Soong *et al.* 2010). In this study, we show that considering genetic information, under the form of familial relationships, SNP-derived PCs or WGP using markers evenly distributed in the genome can increase the prediction accuracy of risk of developing skin cancer.

Simply considering family history, under the form of pedigree relationships linked to phenotypes, increased prediction accuracy. However, there is a limit to how much prediction accuracy can be gained by considering family history alone. One of the limitations of using pedigree connections in a predictive model is that family size in humans is usually small. Our sample, for instance, has some unrelated individuals. Other limitations of models using pedigrees are (1) they capture important elements of genetic variability, such as variability due to substructure or admixture, and (2) these models cannot describe genetic differences between individuals with identical pedigree (*e.g.*, full sibs) due to sampling of alleles at meiosis. Therefore, describing genetic background using markers can potentially boost prediction accuracy above and beyond what can be achieved using family history. In agreement with previous studies in animals (VanRaden *et al.* 2009), our study confirms this and suggests that simply considering two-SNP–derived PCs can increase prediction accuracy substantially. Skin color, and therefore ethnicity, is known to be highly correlated with skin cancer (World Cancer Report 2008), and skin color varies even among individuals of European descent. We found that cancer incidence was high at low levels of PC1 and at high levels of PC2. In previous studies, PC1 among Europeans has been found to correspond to ancestry along the northwest to southeast European geographical axis (Campbell *et al.* 2005; Novembre *et al.* 2008; Drineas *et al.* 2010). Therefore, one possible explanation of the increase in prediction accuracy obtained by considering two-SNP–derived PCs is that these PCs are capturing ancestry, which correlates with skin color and with risk of developing skin cancer.

Naturally, there is also a limit in the proportion of genetic variability at causal loci that can be explained by two PCs. Our study confirms this: the WGP model using 41K-SNP outperformed all the other models we considered, including the one with PCs. The prediction accuracy of WGP increases

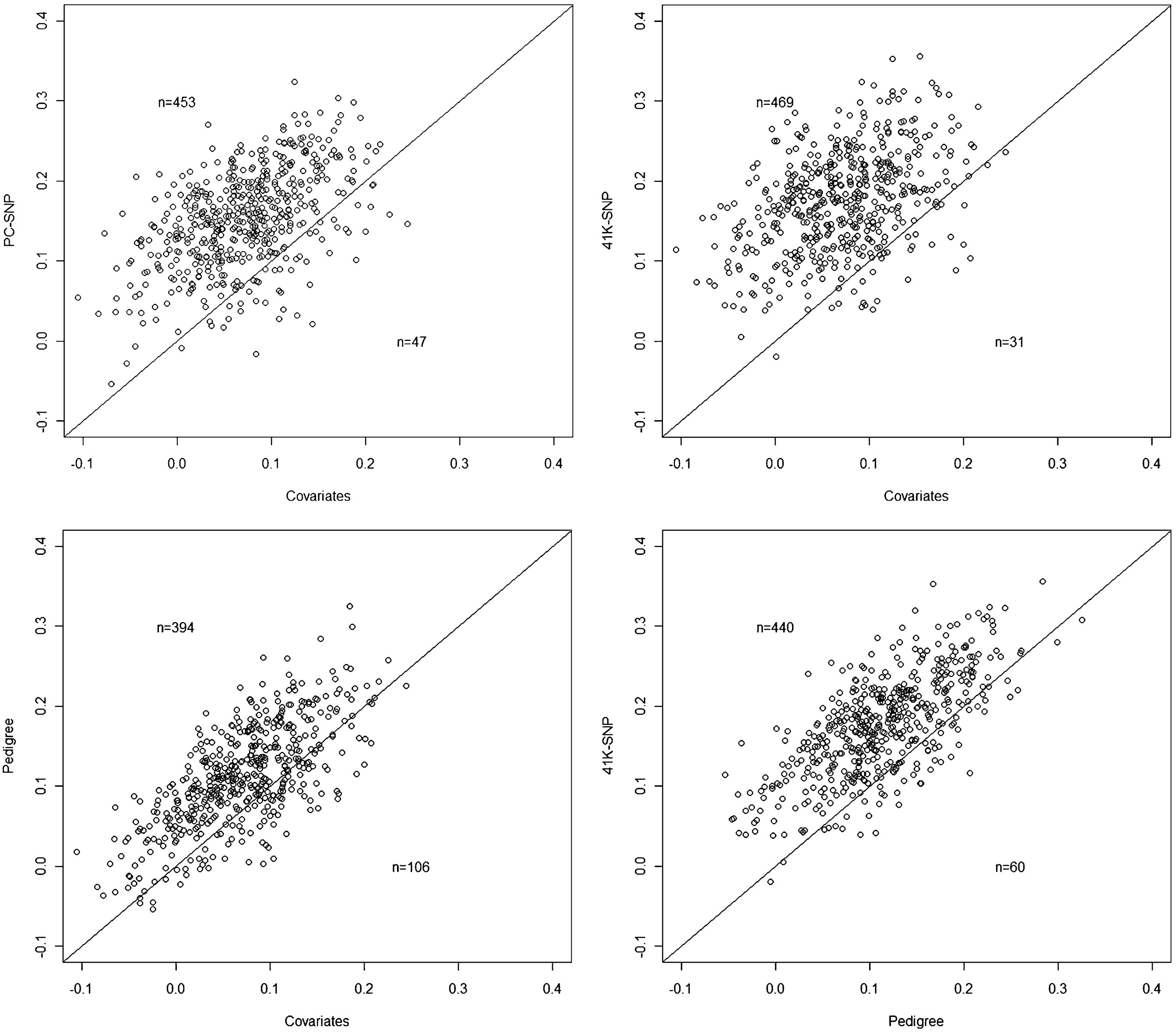

Figure 5 AUC in 500 random training–testing sets of genetically informed models (pedigree model and PC-SNP and 41K-SNP models) *vs.* the baseline model (covariates) and average AUC for the 500 training–testing sets in the 41K-SNP model *vs.* the pedigree model.

monotonically with marker density, a finding that is consistent with the hypothesis that genetic risk to skin cancer is affected by a large number of variants. A polygenic genetic architecture has also been suggested for other human traits (*e.g.*, Vattikuti *et al.* 2012).

Empirical evidence for complex traits in animals (Vazquez *et al.* 2010) and humans (Makowsky *et al.* 2011) has shown that prediction accuracy increases with marker density, and our results are consistent with this. However, prediction accuracy depends on many other factors, perhaps most importantly on the size of the training set (*n*) (VanRaden *et al.* 2009). Preliminary evidence of (unpublished) studies we conducted with human height suggests that the level at which the curve relating prediction accuracy and marker density reaches a plateau is highly dependent on marker density. Prediction accuracy also depends on the selection criteria of the markers incorporated as predictors in the model (Vazquez *et al.* 2010). In that study, the predictive correlation obtained with a set of 300 markers highly associated with the trait of interest was, on average for six traits in cattle, 0.18 higher than that obtained with a set of 300 evenly spaced markers. Here, markers are evenly distributed, representing the whole genome, and are not particularly associated with skin cancer. We expect predictive accuracy to increase if the markers were selected based on their association with the disease. Therefore, we speculate that further increases in marker density, targeting markers associated with skin cancer, accompanied by increases in sample size, could increase prediction accuracy of WGP even further.

The WGR models implemented in this study account for additive effects of markers. Potentially, these additive models could be extended to account for interactions of alleles within loci (*i.e.*, dominance) and between loci (*i.e.*, epitasis). With *p* markers, modeling additive and dominance effects involves estimating 2*p* effects, and this can be done by using the methods similar to those described in this article. However, modeling epistatic interactions is much more difficult because the number of contrasts required and, consequently, the number of parameters to be estimated grow exponentially with the number of markers and the order of the interaction. Alternatively, one can attempt to capture departures from the linear model, using WGP with nonparametric procedures, such as penalized neural networks or reproducing kernel Hilbert spaces (Gianola *et al.* 2006; de los Campos *et al.* 2010b). However, even in cases where complex interactions among alleles hold at the causal level, a large proportion of interindividual differences in genetic risk may manifest as additive variance (Hill *et al.* 2008), and the information provided by data for estimation of nonadditive effects may be small (Hill *et al.* 2008). Because of this, it is not necessarily the case that use of models that account for nonadditive effects will yield higher prediction accuracy than that of an additive model.

Summary

Although accurate in predicting survival rate once the signs of the disease are present, previous predictive models for skin cancer do not account for genetic susceptibility factors (Soong *et al.* 2010), and therefore they have limited use for preventive measures that can be applied early in life. In our study, prediction substantially improved by using genetic parameters in the predictive models. Further, methods including genome-wide markers information outperformed models with genetic risk estimates derived from the pedigree. WGP is a promising tool for estimating individual genetic predisposition to skin cancer before it is detected or even developed. We speculate that genomic information may be used to prospectively identify individuals with particularly high risk of developing skin cancer.

Acknowledgments

We gratefully acknowledge scientists conducting the Framingham Heart Study for collecting, analyzing, and providing the genotypic and phenotypic data used in the present study. We also acknowledge the invaluable dedication of all participants in the Framingham Heart Study for their contribution over many years. We acknowledge National Institute of Health grants R01GM077490, R01GM099992, and R01GM101219, and a Kraft grant, as well as the Wisconsin Agriculture Experiment Station, for their financial support.

Literature Cited

Bulmer, M. G., 1980 *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, New York.

Campbell, C. D., E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman *et al.*, 2005 Demonstrating stratification in a European American population. Nat. Genet. 37: 868–872.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, and J. Burgueño *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713.

Dawber, T. R., G. F. Meadors, and F. E. Moore Jr., 1951 Epidemiological approaches to heart disease: the Framingham Study. Am. J. Public Health 41: 279.

Dawber, T. R., W. B. Kannel, and L. P. Lyell, 1963 An approach to longitudinal studies in a community: the Framingham Study. Ann. N. Y. Acad. Sci. 107: 539–556.

de los Campos, G., and P. Pérez, 2010 *BLR: Bayesian Linear Regression. R package version 1.2* Manual available at: http://CRAN.R-project.org/package=BLR. Accessed: October 22, 2012.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375.

de los Campos, G., D. Gianola, and D. B. Allison, 2010a Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. 11: 880–886.

de los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa, 2010b Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92: 295–308.

Dempster, E. R., and I. M. Lerner, 1950 Heritability of threshold characters. Genetics 35: 212.

Diepgen, T., and V. Mahler, 2002 The epidemiology of skin cancer. Br. J. Dermatol. 146: 1–6.

Drineas, P., J. Lewis, and P. Paschou, 2010 Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. PLoS ONE 5: e11892.

Fawcett, T., 2006 An introduction to ROC analysis. Pattern Recognit. Lett. 27: 861–874.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 52: 399–433.

Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.

Goddard, M. E., and B. J. Hayes, 2007 Genomic selection. J. Anim. Breed. Genet. 124: 323–330.

Gudbjartsson, D. F., P. Sulem, S. N. Stacey, A. M. Goldstein, and T. Rafnar *et al.*, 2008 ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. Nat. Genet. 40: 886–891.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Han, J., P. Kraft, G. A. Colditz, J. Wong, and D. J. Hunter, 2006 Melanocortin 1 receptor variants and skin cancer risk. Int. J. Cancer 119: 1976–1984.

Harville, D. A., and R. W. Mee, 1984 A mixed-model procedure for analyzing ordered categorical data. Biometrics 40: 393–408.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009 Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Hill, W. G., 2010 Understanding and using quantitative genetic variation. Philos. Trans. R. Soc. B Biol. Sci. 365: 73.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4: e1000008.

IARC, 1992 Solar and ultraviolet radiation. IARC Monogr. Eval. Carcinog. Risks Hum. Lyon, France. 55: 1–316.

International Agency for Research on Cancer, World Health Organization, 2008 *World Cancer Report 2008*, edited by P. Boyle and B. Levin, World Health Organization, Lyon, France. Available at http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/index.php. Accessed August 1, 2011.

Kreger, B. E., G. L. Splansky, and A. Schatzkin, 1991 The cancer experience in the Framingham Heart Study cohort. Cancer 67: 1–6.

Lander, E. S., and N. J. Schork, 1994 Genetic dissection of complex traits. Science 265: 2037.

Lee, S. H., J. H. J. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher, 2008 Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. 4: e1000231.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte *et al.*, 2011 Beyond missing heritability: prediction of complex traits. PLoS Genet. 7: e1002051.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, and L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461: 747–753.

McCarthy, E. M., K. P. Ethridge, and R. F. Wagner Jr., 1999 Beach holiday sunburn: the sunscreen paradox and gender differences. Cutis 64: 37–42.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature 456: 98–101.

Park, T., and G. Casella, 2008 The Bayesian lasso. J. Am. Stat. Assoc. 103: 681–686.

Pérez, P., and G. de los Campos, J. Crossa, and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome J. 3: 106.

Pérez-Cabal, M. A., A. I. Vazquez, D. Gianola, G. J. Rosa, and K. A. Weigel, 2012 Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. Front Genet. 3: 27.

Pharoah, P. D., 2008 Shedding light on skin cancer. Nat. Genet. 40: 817–818.

Price, A. L., J. Butler, N. Patterson, C. Capelli, and V. L. Pascali *et al.*, 2008 Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 4: e236.

Purcell, S. M., N. R. Wray, J. L. Stone, P. M. Visscher, and M. C. O'Donovan *et al.*, 2009 Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752.

R Development Core Team, 2010 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Robinson, J. K., 1990 Behavior modification obtained by sun protection education coupled with removal of a skin cancer. Arch. Dermatol. 126: 477.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, 2005 ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940.

Soong, S., S. Ding, D. Coit, C. M. Balch, J. E. Gershenwald *et al.*, 2010 Predicting survival outcome of localized melanoma: an electronic prediction tool based on the AJCC Melanoma Database. Ann. Surg. Oncol. 17: 2006–2014.

Stacey, S. N., D. F. Gudbjartsson, P. Sulem, J. T. Bergthorsson, and R. Kumar *et al.*, 2008 Common variants on 1p36 and 1q42 are associated with cutaneous basal cell carcinoma but not with melanoma or pigmentation traits. Nat. Genet. 40: 1313–1318.

Stern, R. S., 2010 Prevalence of a history of skin cancer in 2007: results of an incidence-based model. Arch. Dermatol. 146: 279.

Thomas-Ahner, J. M., B. C. Wulff, K. L. Tober, D. F. Kusewitt, J. A. Riggenbach *et al.*, 2007 Gender differences in UVB-induced skin carcinogenesis, inflammation, and DNA damage. Cancer Res. 67: 3468–3474.

Tian, C., R. M. Plenge, M. Ransom, A. Lee, and P. Villoslada *et al.*, 2008 Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet. 4: e4.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.*, 2009 Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16–24.

Vattikuti, S., J. Guo, and C. C. Chow, 2012 Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet. 8: e1002637.

Vazquez, A. I., 2010 Statistical modeling of genomic data: applications to genetic markers and gene expression. Ph.D. Dissertation, University of Wisconsin, Madison, WI.

Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola *et al.*, 2010 Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93: 5942–5949.

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola *et al.*, 2010 Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93: 5423–5435.

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2008 Prediction of individual genetic risk of complex disease. Curr. Opin. Genet. Dev. 18: 257–263.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569.

Ziegler, A., A. S. Jonason, D. J. Leffellt, J. A. Simon, H. W. Sharma *et al.*, 1994 Sunburn and p53 in the onset of skin cancer. Nature 372: 773–776.

Communicating editor: F. Zou